International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Survey on Speech Emotion Recognition System Using CNN Algorithm

G. Prasanth Kumar[1], K. Anisha[2], K. Vamsi Priya[3], Ch. Sahithi[4], N. Devi Charan[5], A.V.S. SivaRamaRao[6]

[1,2,3,4,5,6]Department of Computer Science and Engineering, Sasi Institute of Technology & Engineering

Abstract

The speech emotion recognition system (SER) plays an important role in decoding and predicting the speaker's emotional state by analyzing audio signals. Emotions are often simplified and grouped into categories such as anger, happiness, sadness, and even neutral emotional states. These emotions serve to communicate one's thoughts and provide insight into one's physical and mental health. Speech is the most basic and natural form of communication when interacting with others. Voice signals enable rapid communication between individuals, making them a valuable and effective method of expression. Over the past decades, countless research initiatives have been devoted to the development of voice-based automatic emotion recognition systems, especially to improve human-machine communication. Voice is gradually taking center stage in the field of Human-Machine interfaces in the IT sector. This interdisciplinary field draws on a variety of fields, including computer science, signal processing, psychology, linguistics, and more. As technology advances, it allows for seamless communication between humans and machines. Speech recognition not only interprets speech but also captures nuances in a person's tone and expressions, similar to body language. Therefore, it becomes an essential element of the Human-Machine communication system.

Keywords: Speech Emotion Recognition (SER), CNN Algorithm, and Artificial Intelligence, Dataset.

I. INTRODUCTION

Human life is profoundly intertwined with emotions, playing a crucial role in our daily tasks and shaping our perceptions of the world. Emotions are conveyed by a variety of means, including speech, facial expressions, and gestures. Recognize speech emotions is the study of analyzing vocal activity as a key component of understanding and interpreting these emotions.

This concept is supported by substantial evidence indicating that many emotional states expressed during speech are accompanied by physiological changes that impact the way a person modulates their voice. For example, when someone experiences anger, it often leads to alterations in their breathing patterns, heightened muscle tension, and a shift in the geometry of their vocal tract. These physiological changes also affect the vibration of the vocal folds and the acoustic properties of their speech. In essence, emotions are intricately connected to the way we speak, and recognizing these emotional cues in speech can provide valuable insights into human communication and behaviour.

The most predominant mode of human communication is speech, a rich medium that encompasses an abundance of paralinguistic information. Within speech, elements like Body language, gestures, facial

expressions, voice and tone can all play a role in quickly conveying information, allowing individuals to obtain immediate information about characteristics such as age, gender calculation, etc. Speech and emotion recognition (SER) is an exciting area of computer science research that has seen significant growth over the past few decades. Its applications span a variety of sectors, including smart home automation, social media, education, healthcare, and many AI-based applications, and provide framework simple to automatically identify emotions in speech. Innovative strategies in SER research include using adversarial training techniques to generate diverse and realistic speech data, thereby expanding the training dataset and improving system performance SER[1,2]. This innovative approach promises to further enhance the accuracy and reliability of automated RES, opening up exciting possibilities for better human-machine interaction and understanding.

Creating effective features for recognizing emotions in speech is a challenging aspect of Speech Emotion Recognition (SER) because features derived from unprocessed speech signals can successfully distinguish between different emotional states. SER is a technology-driven process that leverages cutting-edge techniques to identify human emotions from recorded speeches or real-time situations [3]. This not only showcases the remarkable advancements achieved in the field of machine learning but also holds significant implications for a wide range of industries, including entertainment, healthcare, and customer service.

By enabling machines to comprehend and respond to human emotions, we are setting the foundation for more compassionate and contextually aware AI systems. Enhancing our ability to engage with machines on a deeper emotional level and opening up new possibilities for human-computer interaction [4]. In this comprehensive survey study, we will delve into the key approaches, datasets, and state-of-the-art techniques employed in the field of speech emotion recognition. Our ultimate aim is to unravel how machines can decipher the intricate tapestry of human emotions embedded within spoken words. Along this journey, we will explore everything from the fundamental principles of acoustic feature extraction to the latest advancements in deep learning architectures. By teaching machines to recognize and respond to human emotions, we are paving the way for a future characterized by more empathetic and context-aware AI systems.

Emotions can be conveyed through various means, including facial expressions, tone of voice, and body language, and this capability holds implications not only for emotion recognition but also for related fields such as emotional computing and human-computer interaction [5-6].

The development of an ideal Speech Emotion Recognition framework involves precise calculations and the utilization of specific datasets to train the machine or system to identify and categorize these emotions based on the words used or the tone of voice. Bridging the gap between acoustic features (related to sound intensity and frequency patterns) and human emotions (e.g., happiness, sadness) makes automated Speech Emotion Recognition a complex endeavor , heavily reliant on capturing discernible acoustic features from a given recognition task.[7]As we navigate this landscape, we will also delve into the ethical considerations surrounding emotion recognition technology, its potential societal impacts, and the vital importance of responsible development and deployment.

Furthermore, we will explore real-world applications where speech emotion recognition is already making a positive difference, ranging from mental health support systems to virtual assistants that adapt to users' emotional states. This research not only enhances our understanding of the field but also sheds light on its broader implications for society and human interaction.[8]

The section II is about the Background study and section III speaks about related work, section IV it brief about analysis and discussion of section III, section V addresses the challenges involved in encrypting an image and finally section VI provides the conclusion of the paper.

II. BACKGROUND STUDY

Speech Emotion Recognition (SER) has emerged as a dynamic field of research and development, driven by the growing need to improve human-computer interaction and communication systems. Understanding and interpreting emotions conveyed through speech is crucial for creating more empathetic and responsive technology. SER systems are designed to automatically detect, analyze, and categorize emotions expressed by individuals during spoken communication. This technology has found applications in various domains, including customer service, mental health assessment, human-robot interaction, and beyond.

The foundation of SER lies in the rich body of research in fields such as speech processing, signal analysis, machine learning, and psychology. Researchers have been exploring the intricate relationship between acoustic features of speech, prosody, and the underlying emotional state of speakers. These investigations have led to the development of sophisticated algorithms and models that can recognize emotional cues from speech data. As SER technology advances, it not only benefits from more extensive and diverse datasets but also from the integration of deep learning techniques, enabling more accurate and nuanced emotion recognition.

The importance of SER extends beyond improving technology interfaces. It has applications in healthcare, where it can aid in diagnosing and monitoring mental health conditions by analyzing speech patterns for signs of depression, anxiety, or other emotional states. In education, SER can be employed to assess student engagement and tailor instructional content accordingly. Furthermore, as the Internet of Things (IoT) and smart devices become more prevalent, SER can enhance the adaptability and responsiveness of these systems to better serve users' emotional needs.

Challenges in SER research include cross-cultural variations in emotional expression, the need for robust real-time recognition, and ethical considerations related to privacy and consent. As SER technology matures, addressing these challenges will be essential for its widespread adoption. Moreover, with the ever-increasing integration of speech and emotion recognition into our daily lives, SER is poised to play a pivotal role in shaping the future of human-computer interaction, making technology more intuitive, empathetic, and aligned with human emotional experiences.

III. RELATED WORK

In this section, briefly explain about the speech emotion recognition using various deep learning methods by researchers.

The core of the Speech emotion recognization (SER) system lies in the CNN (Convolutional Neural Network) algorithm, which comprises modules for both emotion detection and emotion classification. These modules play a crucial role in distinguishing between various emotions, including happy, surprise, anger, neutrality and sad. The dataset consists of speech samples.

The Author Apoorv Singh *et al*. [9], proposed a model that incorporates the CNN algorithm for SER and integrates it with robots and music applications. The SER system will better understand the mood of the corresponding human by helping the system to converse with the human more effectively. In addition, the proposed system can be integrated with Music apps, so as to recommend songs to users based on their

mood or emotions. For emotion distinction tasks, they created many models and identified the best CNN model. The new model was 71% accurate, but could have done better if it could distinguish voices of male and female. The Integrating CNN-based SER into robots and music apps enhances user experiences by enabling emotional awareness and adaptive responses.

The Researcher Huihui *et al*. [10] introduce an innovative approach to improve the performance of traditional Convolutional Neural Networks enhancing discriminative power in emotion-related features. Their proposed model is called ICNN. The key advancement in the ICNN model involves the use of interactive Convolution processes for feature maps of varying scales, resulting in a notable achievement of 76% accuracy in emotion recognition. Specifically, they partition the Mel-frequency cepstral coefficients (MFCs) into parallel channels, denoted as H-MFC and L-MFC, using the ICNN process. Integrating CNNs with MFCC or LMFCC features can leverage the strengths of both feature engineering and deep learning, leading to improved performance and robustness in audio and speech-related applications.

The Author Taiba *et al*. [11] have focused on the identification of emotions and the development of robust methods for their detection. Innovative advancements in the field of Speech Emotion Recognition (SER) have been introduced through the development of a novel approach known as Deep Stride Convolutional Neural Networks (DSCNN). This pioneering modification revolves around the adaptation of Convolutional Neural Networks (CNNs), specifically focusing on the elimination of pooling layers while prioritizing the utilization of strides to effectively reduce the dimensionality of feature maps. To evaluate and compare the effectiveness of the traditional CNN model against the newly proposed DSCNN model, a research paper conducted two comprehensive experiments. In both experiments, the performances of the models were improved as more epochs were added to the training process. DSCNN performance was better than the state of-the-art CNN model, with 87.8% accuracy obtained by DSCNN compared to 79.4% by CNN. This refers that for convincingly detecting emotions, CNN requires further enhancement in the architecture.

The Researcher Ming-Hsiang Su *et al* [12]. novel approach to discourse feeling acknowledgment that considers both verbal and nonverbal sounds inside an expression and gotten 68.87% execution. The proposed framework employments a back vector machine-based finder and utilizes profound remaining systems and a mindful long short-term memory-based sequence-to-sequence show to attain noteworthy comes about on the NNIME corpus. The sound type features and nonverbal vocalization was helpful for emotion recognition and improved the accuracy of their proposed method. This method is also used in robust to high dimensional data and helps in reducing the risk of overfitting by introducing the skip connections.

Chenghao Zhang *et al* [13]. propose a novel calculation that viably extricates emotion-oriented highlights to move forward the execution of the Discourse Feeling Acknowledgment (SER) framework. The proposed calculation is called Autoencoder with Feeling Implanting, which can proficiently learn a priori information from the name and recognize which highlights are most related to human feeling. To improve the performance which is of 65.76% of the SER system by extracting emotion features and making use of data effectively. An autoencoder with emotion embedding is a neural network architecture that combines auto encoders with the capability to learn and represent emotional information.

The Researcher Jorge Oliveira *et al* [14]. investigate the utilization of pre-trained discourse acknowledgment profound layers to identify feelings within the working environment, with the objective of optimizing fabricating forms and expanding efficiency by killing or weakening negative feelings which

gotten 89.43% of precision and this paper included preparing and testing a calculation on the Merge dataset, which is an English multiparty dataset collected from the TV appear "Friends". A weighted strategy was implemented, resulting in pre-trained speech recognition layers serve as a valuable starting point for developing accurate and efficient speech processing systems across a range of applications, Transfer Learning and Reduced Data Requirements.

Julia Sidorova *et al*. [15] conducted research with the aim of presenting a framework for the automated evaluation of emotional competence in neurological patients, particularly those with Foreign Accent Syndrome (FAS) and obtained 87.89% of accuracy. The Aggregated Ear model to draw conclusions about the level of competence demonstrated by patients. They also provide a complete description of each individual therapy and its expected neurological effect. Aggregated Ear Models, also known as Ensembled Ear Models, refer to the practice of combining multiple machine learning models or neural networks to improve performance, particularly in the context of ear-related tasks such as ear recognition or biometrics. It's important to note that the effectiveness of aggregated ear models depends on the specific task, the quality and diversity of the base models, and the ensemble technique used (e.g., bagging, boosting, stacking) Tackling Imbalanced Data and Real-world Applicability.

S. Hamsa *et al* [16] aimed to pioneer the development and implementation of an artificial emotional intelligence system that possesses the remarkable ability to accurately identify an unknown speaker's emotional state even in the presence of disruptive noise and interference, achieving an impressive accuracy rate of 90.76%. Their innovative approach hinges on a novel framework for emotion recognition that takes into account critical factors, namely energy, temporal dynamics, and spectral features. To achieve their goals, the researchers harnessed the power of a random forest classifier in conjunction with a wavelet packet transform-based cochlear filter bank (WPT-CFB) combined with a Random Forest classifier offers several advantages in various audio and speech processing tasks. The WPT-CFB and Random Forest combination offers several advantages, it's essential to consider the nature of the data, and the availability of label samples for training.

Srinivas Parthasarathy *et al*. [17] aim to enhance the generalization and performance of models by incorporating unsupervised auxiliary tasks and leveraging both labeled and unlabeled data. They propose and evaluate a semi-supervised approach for speech emotion recognition using ladder networks, achieving a performance rate of 65.76%. They assess the approach's performance using various feature inputs and cross-corpus scenarios, comparing it to fully supervised and multitask learning baselines and evaluate its performance on different feature inputs and cross-corpus scenarios as well It's important to consider factors such as dataset quality, model architecture, and real-world deployment challenges when developing real-time SER regarding user privacy and consent.

Sudarsana Reddy Kadiri *et al*. [18], the goal is to automatically identify emotions in speech using excitation parameters derived around glottal closure instants. They utilize features like instantaneous fundamental frequency, achieving a performance of 86.65%. These excitation features are computed using signal processing techniques, including zero-frequency filtering and linear prediction analysis. The benefits of the paper are Excitation features around glottal closure instants of the features capture information related to the opening and closing of the vocal folds during speech production and Improved Speech Processing.

Siddique Latif *et al*. [19] introduce a multi-task learning framework for speech emotion recognition, designed to improve performance, with an achieved accuracy of 65.43%. This framework includes auxiliary tasks and an adversarial autoencoder. The research demonstrates that adding secondary tasks

and additional data can significantly enhance the accuracy of the primary task and the weight of the loss functions for the primary and secondary tasks should be carefully tuned for optimal performance and the advantages of the multi-task learning framework make it a powerful tool for improving model performance, robustness, and efficiency in various machine learning and deep learning applications.

In the study conducted by Chang Li *et al.* [20], a technique for robotic emotion recognition is introduced, employing two-level feature fusion in speech audio signals. This approach utilizes two layers of BILSTM, feature extractors like VGGISH and MFCC, and fully connected networks for categorization. The research demonstrates that by varying the shift step length of the VGGISH feature extractor and fusing various features, the model achieves emotion recognition accuracy of 69.3% on the IEMOCAP dataset, surpassing the current state-of-the-art model by 0.5% in emotion recognition based on speech. They provide more operable way to enhance insufficient data and improve the performance of robotic emotion recognition systems and the benefits are Two-level feature fusion is a versatile and powerful technique that can be applied in a wide range of domains and applications, enhancing model performance, robustness, and interpretability.

Zhichao Peng *et al*. [21] propose a speech emotion identification system that combines an attention-based backend with a front end based on aural perception. The system introduces an ASRNN (Attention-based Speech RNN) to continuously scan the temporal sequence and focus on emotional regions. It employs a 3D convolution model to capture both local features and periodicity information of emotional speech by jointly learning spectral and temporal features from modulation cues, achieving an accuracy rate of 85.46%. The advantages are auditory perception-based front-end with an attention-based back-end can lead to more accurate, robust, and context-aware speech recognition systems, which are essential for various applications in speech and audio processing including virtual assistants, voice-controlled devices.

Huan Zhao *et al*. [22], a strategy for speech emotion recognition (SER) is discussed, and prospective future research directions are outlined. The strategy's success is highlighted as the SSGAN (Semi-Supervised Generative Adversarial Network) and VSSSGAN (Variational Semi-Supervised Generative Adversarial Network) methods outperform state-of-the-art methods in both intra- and inter-domain contexts. They also suggest that future research to increase the accuracy which is already given as 61.25% could explore the use of other types of generative models or investigate the impact of different types of unlabelled data on the performance of the SSGAN. Overall, the proposed approach shows promise for improving the performance of SER systems by leveraging both labelled and unlabelled data.

Joseph Bamidele Awotunde *et al.* [23] propose a method for enhancing speech understanding in noisy environments using speech segregation via a convolutional neural network (CNN). Their findings indicate that the proposed method outperforms some contemporary speech processing techniques by 78.52% in specific input SNR (Signal-to-Noise Ratio) settings. The authors suggest that their method can be practically applied to enhance voice recognition technologies in noisy settings. This method incorporates convolutional neural networks (CNNs) and deep learning models, showcasing their effectiveness in various speech processing tasks, including speech segregation.

Ting-Wei-Sun *et al* [24]. using deep learning and including more data, including speaker gender, can enhance the precision of voice emotion recognition algorithms. The experimental results demonstrate that, in comparison to existing speech emotion detection algorithms in various language systems, the proposed approach generated much higher-accuracy predictions. improved speech processing, enhanced personalization, customer service and calls, and feedback were all advantages of this paradigm.

The Researcher Guanglong Du *et al* [25]. a technique for non-contact emotion recognition that makes use of video-based heart rate and facial expression data, and they assess how it may be used to improve interactive gaming settings. Additionally, the research tries to overcome the shortcomings of existing emotion detection techniques that concentrate on discrete signals and are unable to distinguish between different emotional intensities. They include the development of a method for non-contact emotion recognition with result of 67.43% accuracy using video basing detection of heart rate and facial expression features, and the use of a SOM-BP network for real-time processing the model used in this paper is best for Non-Invasiveness, Real-Time Analysis, Scalability, Improved Customer Service, COVID-19 Adaptability.

In the research conducted by Qiuqiang Kong *et al*. [26], employ neural networks pretrained on extensive datasets to address challenges related to audio pattern identification. They report an impressive accuracy rate of 79.54% in their approach. Their work has significant potential applications in real-world scenarios, including audio tagging, acoustic scene classification, and sound event identification. Notably, their proposed approach surpasses previous systems developed on specialized datasets with limited durations.

Zhen Tao Liu *et al* [27]. In order to decrease the number of characteristics needed to train models and increase accuracy, they suggested a speech personality recognition with essential audio features model. This model's accuracy is 76.43%. The suggested method uses loglikelihood distance for annotation categorization and audio feature extraction. The outcomes demonstrate that this strategy outperforms current approaches and has potential for usage in practical situations.

Rajdeep Chatterjee *et al*. [28], the authors present a strategy suitable for integration into consumer electronics devices for household use. Their research highlights the effectiveness of a 1-D Convolutional Neural Network (CNN) for Speech Emotion Recognition (SER), particularly when considering real-time operation and recognition performance. The proposed model's robustness is compared with other state-of-the-art techniques, and empirical observations are obtained with 98.65% accuracy of this model are Efficient Feature Extraction and Real-time Processing.

Mauajama Firdaus *et* al. [29], have introduced a pioneering endeavor focused on sentiment and emotion-controlled dialogue generation in a multi-modal context. Their research investigates the collective influence of sentiment and emotion in shaping the process of generating dialogues. They employ a diverse range of data sources, encompassing text, video, audio, reinforcement learning, and user feedback, as part of their strategy to augment system performance. Impressively, their model attains a commendable accuracy rate of 89.76%. This innovative approach carries substantial potential benefits across a spectrum of applications.

Mohammad Ariff Rashidan *et al*. [30] proposes a model for affective states in consumer electronics voice analysis. The authors acknowledge certain limitations in terms of scope, chosen databases, time period, and inclusion/exclusion criteria. Their work focuses on emotion recognition in technology-assisted communication with autistic children.

IV. ANALYSIS AND DISCUSSIONS

Speech Emotion Recognition (SER) is a field of research that has witnessed remarkable progress over the past decade. With the proliferation of voice-activated devices and growing interest in sentiment analysis, the demand for effective SER systems has soared.

Speech Emotion Recognition (SER) has gained significant attention in recent years due to its wide range of applications in fields such as human-computer interaction, virtual assistants, sentiment analysis, and

mental health assessment. This technology involves the automatic identification and classification of emotional states from spoken language.

TABLE I. ANALYSIS OF COMMON METRIC

Authors	Model/Algorithm	Accuracy
Apoorv *et al* [9] (2020)	CNN algorithm for SER with integration into robots and music apps	71%
Huihui *et al.* [10] (2020)	ICNN with MFCC and LMFCC	76%
Taiba [11] *et al.* (2020)	Modified CNN model	87.8%
Ming *et al.* [12] (2021)	Support Vector Machine-based detector, deep residual networks, and attentive LSTM-based sequence-to sequence model	68.87%
C. Zhang *et al.* [13] (2021)	Autoencoder with Emotion Embedding	65.76%
Julia *et al.* [15] (2021)	Aggregated Ear model	87.89%
S. Hamsa *et al.* [16] (2020)	Random forest classifier	90.76%
Parthasarathy *et al.* [17] (2020)	Semi - supervised ladder networks	65.76%
Sudarsana *et al.* [18] (2020)	excitation features around glottal closure instants	86.65%
Siddique *et al.* [19] (2020)	Multi-task learning framework	65.43%
Chang *et al.* [20] (2022)	Two-level feature fusion	69.3%
Peng *et al.* [21] (2020)	Front-end based on auditory perception, back-end based on attentiveness	85.46%
Zhao *et al.* [22] (2020)	SSGAN and VSSSGAN methods	61.25%
Awotunde *et al.* [23] (2020)	CNN based method for speech segregation	78.52%
Wei sun *et al.* [24] (2020)	Deep learning with speaker gender	95.65%
Guang *et al.* [25] (2020)	Non-Contact emotion recognition using videobased features	67.43%
Qiuqiang *et al.* [26] (2020)	Neural networks pretrained on large-scale datasets	79.54%
Zhen *et al.* [27] (2021)	Speech personality recognition with essential audio features	76.43%
Mauajama *et al.* [29] (2022)	Sentiment and emotion-controlled dialogue	89.76%
Mohammad *et al.* [30] (2021)	Affective states in speech analysis for consumer Electronics	97.5

A. Autocorrelation

The autocorrelation of the speech signal at different lag values and then find the lag that corresponds to the highest peak in the autocorrelation function.

The formula for autocorrelation at a lag value (k) is:

$$R(k) = \sum_{n-0}^{N-K-1} x(n) \, . \, x(n+k)$$

B. RMSE

RMSE (Root Mean Square Error) can be used when predicting continuous emotion values. It measures the average squared difference between predicted and actual continuous emotion values.

$$RMSE = \sqrt{\frac{1}{N}} \sum_{i=1}^{N} (Predicted_i - Actual_i)^2$$

CHALLENGES AND GAPS

Emotion Variability: Emotions are expressed differently by individuals, and cultural, regional, or individual factors can influence emotional expression. The challenge is to build models that can accurately recognize and classify a wide range of emotions while accounting for this variability.

Speaker Variability: Speech recognition systems must handle speaker variability, including differences in pitch, accent, and speaking style. These differences can influence the way emotions are conveyed and create challenges in building models that generalize effectively across various speakers.

Privacy and Ethical Concerns: As SER technology advances, concerns about privacy and ethical use have become more pronounced. Gaps in guidelines and regulations for the responsible and ethical deployment of SER systems need to be addressed.

Robustness to Noise and Environmental Factors: Building SER systems that are robust to various noise sources and environmental conditions is a persistent gap. Achieving reliable performance in real-world, noisy settings is a priority**.**

Contextual Information: Understanding emotions often requires context beyond audio data. Integrating textual transcripts or visual cues (such as facial expressions) into the recognition process can enhance accuracy but also introduces challenges related to multimodal data fusion.

Interdisciplinary Collaboration: Bridging the gap between speech processing and emotion psychology is vital. Collaborative research between experts in both fields can lead to a deeper understanding of the emotional cues present in speech.

The design and implementation of SER algorithms are crucial in addressing these challenges. Researchers continue to work on developing new and improved SER algorithms that address these challenges and provide robust and secure image encryption solutions. These algorithms are designed to handle the diversity in emotional expression, mitigate the impact of noise and environmental factors, and provide accurate emotion recognition across a variety of speakers and languages. Furthermore, they play a role in achieving robust generalization and efficient transfer learning, thus making SER systems more adaptable to diverse contexts.

V. CONCLUSION

This paper primarily focuses on the application of advanced technologies in Discourse-based emotion detection. After reviewing several research papers, we can outline the potential artificial intelligence/machine learning algorithms In the realm of speech emotion detection, various methodologies have been explored, some of which include Convolutional Neural Networks (CNN), Long Short-Term

Memory (LSTM), Deep Convolutional Neural Networks (DCNN), Vector Space Models (VSM), and hybrid approaches. Notably, when it comes to discourse emotion detection and feature extraction, the Convolutional Neural Network (CNN) algorithm and its derivatives, including DCNN, have demonstrated superior accuracy compared to other techniques. These classifiers are pivotal in discerning discrete human emotions such as happiness, anger, neutrality, and more, leveraging the distinctive characteristics present in voice and speech signals.

The success of these approaches has been realized across a wide spectrum of datasets encompassing diverse speech samples. As a result, they hold significant potential in numerous real-world applications, spanning fields like education, healthcare, business process outsourcing (BPO), and crime detection. Recognizing the versatility and impact of Speech Emotion Recognition (SER) systems, there's a growing interest in developing conversational Human-Interaction models that are customized to meet the unique needs and preferences of individual users. These conversational models aim to create more personalized and adaptive interactions between users and machines, offering enhanced user experiences across various domains. By harnessing the power of SER and related technologies, these models can facilitate more empathetic and responsive human-machine interactions, ultimately leading to improved communication and user satisfaction. This research and development in SER and conversational Human-Interaction models represent an exciting frontier with significant implications for the future of technology and human-computer interfaces.

REFERENCES

1. R. R. Sehgal, S. Agarwal and G. Raj, "Intelligent Voice Reaction involving Opinion Examination in Programmed Discourse Acknowledgment Frameworks," 2018 Global Gathering on Advances in Processing and Correspondence Designing (ICACCE), 2018, pp. 213-218, doi:10.1109/ICACCE.2018.8441741.
2. K. Huang, C. Wu, Q. Hong, M. Su and Y. Zeng, "Discourse Feeling Acknowledgment utilizing Convolutional Brain Organization with Sound Word-based Installing," 2018 eleventh Worldwide Conference on Chinese Communicated in Language Handling (ISCSLP), 2018, pp. 265-269, doi: 10.1109/ISCSLP.2018.8706610.
3. J. Cornejo and H. Pedrini, "Bimodal Feeling Acknowledgment In light of Sound and Facial Parts Utilizing Profound Convolutional Brain Organizations," 2019 eighteenth IEEE Global Meeting On AI And Applications (ICMLA), 2019, pp. 111-117, doi:10.1109/ICMLA.2019.00026.
4. Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, "Discourse based Feeling Acknowledgment utilizing AI", 2019 third Global Meeting on Processing Systems and Correspondence (ICCMC)
5. Y. Dong, X. Yang, X. Zhao and J. Li, "Bidirectional Convolutional Intermittent Inadequate Organization (BCRSN): An Effective Model for Music Feeling Acknowledgment," in IEEE Exchanges on Mixed media, vol. 21, no. 12, pp. 3150-3163, Dec. 2019, doi: 10.1109/TMM.2019.2918739.
6. Z. Zhao et al., " Investigating Profound Range Portrayals through Consideration Based Repetitive and Convolutional Brain Organizations for Discourse Feeling Acknowledgment," in IEEE Access, vol. 7, pp. 97515-97525, 2019, doi: 10.1109/ACCESS.2019.2928625.

7. Shahin, A. B. Nassif and S. Hamsa, "Feeling Acknowledgment Utilizing Crossover Gaussian Combination Model and Profound Brain Organization," in IEEE Access, vol. 7, pp. 26777-26787, 2019, doi: 10.1109/ACCESS.2019.2901352.
8. Y. Dong, X. Yang, X. Zhao and J. Li, "Bidirectional Convolutional Intermittent Inadequate Organization (BCRSN): An Effective Model for Music Feeling Acknowledgment," in IEEE Exchanges on Mixed media, vol. 21, no. 12, pp. 3150-3163, Dec. 2019, doi: 10.1109/TMM.2019.2918739.
9. Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan, "Discourse Feeling Acknowledgment Utilizing Convolutional Brain Organization (CNN)", Global Diary of Psychosocial Restoration, Vol. 24, Issue 08 2020.
10. H. Cheng and X. Tang, "Discourse Feeling Acknowledgment in light of Intuitive Convolutional Brain Organization," 2020 IEEE third Worldwide Gathering on Data Correspondence and Sign Handling (ICICSP),2020 ,pp.163167,doi:10.1109/ICICSP50920.2020.9232071.
11. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor and N. Ismail, "Discourse Feeling Acknowledgment utilizing Convolution Brain Organizations and Profound Step Convolutional Brain Organizations," 2020 sixth Worldwide Meeting on Remote and Telematics (ICWT),2020, pp.16, doi:10.1109/ICWT50448.2020.9243622.
12. 2021 File IEEE/ACM Exchanges on Sound, Discourse, and Language Handling Vol. 29," in IEEE/ACM Exchanges on Sound, Discourse, and Language Handling, vol. 29, pp. 3718-3760, 2021, doi: 10.1109/TASLP.2022.3147096.
13. Zhang and L. Xue, "Autoencoder With Feeling Implanting for Discourse Feeling Acknowledgment," in IEEE Access, vol. 9, pp. 51231-51241, 2021, doi: 10.1109/ACCESS.2021.3069818.
14. J.Oliveira and I. Praça, "On the Utilization of Pre-Prepared Discourse Acknowledgment Profound Layers to Recognize Feelings," in IEEE Access, vol. 9, pp. 9699-9705, 2021, doi: 10.1109/ACCESS.2021.3051083.
15. J. Sidorova, S. Karlsson, O. Rosander, M. L. Berthier and I. Moreno-Torres, "Towards Issue Free Programmed Evaluation of Profound Skill in Neurological Patients with a Traditional Feeling Acknowledgment Framework: Application in Unfamiliar Articulation Disorder," in IEEE Exchanges on Emotional Processing, vol. 12, no. 4, pp. 962-973, 1 Oct.- Dec. 2021, doi: 10.1109/TAFFC.2019.2908365.
16. S. Hamsa, I. Shahin, Y. Iraqi and N. Werghi, "Feeling Acknowledgment From Discourse Utilizing Wavelet Parcel Change Cochlear Channel Bank and Irregular Woodland Classifier," in IEEE Access, vol. 8, pp. 96994-97006, 2020, doi: 10.1109/ACCESS.2020.2991811.
17. S. Parthasarathy and C. Busso, "Semi-Directed Discourse Feeling Acknowledgment With Stepping stool Organizations," in IEEE/ACM Exchanges on Sound, Discourse, and Language Handling, vol. 28, pp. 2697-2709, 2020, doi: 10.1109/TASLP.2020.3023632.
18. S. R. Kadiri and P. Alku, "Excitation Elements of Discourse for Speaker-Explicit Feeling Identification," in IEEE Access, vol. 8, pp. 60382-60391, 2020, doi: 10.1109/ACCESS.2020.2982954.
19. S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps and B. W. Schuller, "Perform various tasks Semi-Managed Ill-disposed Autoencoding for Discourse Feeling Acknowledgment," in IEEE Exchanges on Emotional Registering, vol. 13, no. 2, pp. 992-1004, 1 April-June 2022, doi: 10.1109/TAFFC.2020.2983669.

20. Li, "Automated Feeling Acknowledgment Involving Two-Level Elements Combination in Sound Signs of Discourse," in IEEE Sensors Diary, vol. 22, no. 18, pp. 17447-17454, 15 Sept.15, 2022, doi: 10.1109/JSEN.2021.3065012.
21. Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Darn and M. Akagi, "Discourse Feeling Acknowledgment Utilizing 3D Convolutions and Consideration Based Sliding Repetitive Organizations With Hear-able Front-Finishes," in IEEE Access, vol. 8, pp. 16560-16572, 2020, doi: 10.1109/ACCESS.2020.2967791.
22. H. Zhao, Y. Xiao and Z. Zhang, "Powerful Semi directed Generative Antagonistic Organizations for Discourse Feeling Acknowledgment through Circulation Perfection," in IEEE Access, vol. 8, pp. 106889-106900, 2020, doi: 10.1109/ACCESS.2020.3000751.
23. J. B. Awotunde, R. O. Ogundokun, F. E. Ayo and O. E. Matiluko, "Discourse Isolation in Foundation Commotion In view of Profound Learning," in IEEE Access, vol. 8, pp. 169568-169575, 2020, doi: 10.1109/ACCESS.2020.3024077.
24. T. - W. Sun, "Start to finish Discourse Feeling Acknowledgment With Orientation Data," in IEEE Access, vol. 8, pp. 152423-152438, 2020, doi: 10.1109/ACCESS.2020.3017462.
25. G. Du, S. Long and H. Yuan, "Non-Contact Feeling Acknowledgment Consolidating Pulse and Look for Intuitive Gaming Conditions," in IEEE Access, vol. 8, pp. 11896-11906, 2020, doi: 10.1109/ACCESS.2020.2964794
26. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, "PANNs: Enormous Scope Pretrained Sound Brain Organizations for Sound Example Acknowledgment," in IEEE/ACM Exchanges on Sound, Discourse, and Language Handling, vol.28, pp. 2880-2894, 2020, doi: 10.1109/TASLP.2020.3030497.
27. Z. - T. Liu, A. Rehman, M. Wu, W. - H. Cao and M. Hao, "Discourse Character Acknowledgment In view of Explanation Arrangement Utilizing Log-Probability Distance and Extraction of Fundamental Sound Highlights," in IEEE Exchanges on Mixed media, vol. 23, pp. 3414-3426, 2021, doi: 10.1109/TMM.2020.3025108.
28. M. Firdaus, H. Chauhan, A. Ekbal and P. Bhattacharyya, "Emotional Sen: Producing Feeling and Feeling Controlled Reactions in a Multimodal Discourse Framework," in IEEE Exchanges on Emotional Figuring, vol. 13, no. 3, pp. 1555-1566, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.3015491.
29. M. A. Rashidan et al., " Innovation Helped Feeling Acknowledgment for Mental imbalance Range Problem (ASD) Kids: An Orderly Writing Survey," in IEEE Access, vol. 9, pp. 33638-33653, 2021, doi: 10.1109/ACCESS.2021.3060753
30. Y. Huang, J. Xiao, K. Tian, A. Wu and G. Zhang, "Exploration on Vigor of Feeling Acknowledgment Under Ecological Commotion Conditions," in IEEE Access, vol. 7, pp. 142009-142021, 2019, doi: 10.1109/ACCESS.2019.2944386.
31. H. Luo and J. Han, "Nonnegative Lattice Factorization Based Move Subspace Learning for Cross-Corpus Discourse Feeling Acknowledgment," in IEEE/ACM Exchanges on Sound, Discourse, and Language Handling, vol. 28, pp. 2047-2060, 2020, doi: 10.1109/TASLP.2020.3006331.
32. L. Chen, W. Su, M. Wu, W. Pedrycz and K. Hirota, "A Fluffy Profound Brain Organization With Meager Autoencoder for Close to home Expectation Grasping in Human-Robot Communication," in IEEE Exchanges on Fluffy Frameworks, vol. 28, no. 7, pp. 1252-1264, July 2020, doi: 10.1109/TFUZZ.2020.2966167.

33. [33] H. Chen, D. Jiang and H. Sahli, "Transformer Encoder With Multi-Modular Multi-Head Consideration for Nonstop Influence Acknowledgment," in IEEE Exchanges on Media, vol. 23, pp. 4171-4183, 2021, doi: 10.1109/TMM.2020.3037496.
34. [34] L. Yang, D. Jiang and H. Sahli, "Component Enlarging Organizations for Further developing Melancholy Seriousness Assessment From Discourse Signs," in IEEE Access, vol. 8, pp. 24033-24045, 2020, doi: 10.1109/ACCESS.2020.2970496.
35. [35] Y. - P. Ruan and Z. - H. Ling, "Feeling Regularized Contingent Variational Autoencoder for Profound Reaction Age," in IEEE Exchanges on Emotional Figuring, vol. 14, no. 1, pp. 842-848, 1 Jan.- Walk 2023, doi: 10.1109/TAFFC.2021.3073809.
36. P. Chhikara, P. Singh, R. Tekchandani, N. Kumar and M. Guizani, "Combined Learning Meets Human Feelings: A Decentralized System for Human-PC Communication for IoT Applications," in IEEE Web of Things Diary, vol. 8, no. 8, pp. 6949-6962, 15 April15, 2021, doi: 10.1109/JIOT.2020.3037207.
37. Y. - P. Ruan and Z. - H. Ling, "Feeling Regularized Contingent Variational Autoencoder for Profound Reaction Age," in IEEE Exchanges on Emotional Registering, vol. 14, no. 1, pp. 842-848, 1 Jan.- Walk 2023, doi: 10.1109/TAFFC.2021.3073809.
38. Samanta and T. Guha, "Feeling Detecting From Head Movement Catch," in IEEE Sensors Diary, vol. 21, no. 4, pp. 5035-5043, 15 Feb.15, 2021, doi: 10.1109/JSEN.2020.3033431.
39. T. Darn, V. Sethu and E. Ambikairajah, "Pay Procedures for Speaker Changeability in Ceaseless Feeling Expectation," in IEEE Exchanges on Emotional Processing, vol. 12, no. 2, pp. 439-452, 1 April-June 2021, doi: 10.1109/TAFFC.2018.2883044.
40. Z. Zhao et al., " Programmed Evaluation of Wretchedness From Discourse by means of a Progressive Consideration Move Organization and Consideration Autoencoders," in IEEE Diary of Chosen Subjects in Signal Handling, vol. 14, no. 2, pp. 423-434, Feb. 2020, doi: 10.1109/JSTSP.2019.2955012.
41. N. Liu et al., " Move Subspace Learning for Solo Cross-Corpus Discourse Feeling Acknowledgment," in IEEE Access, vol. 9, pp. 95925-95937, 2021, doi: 10.1109/ACCESS.2021.3094355.
42. X. Wu et al., " Model Based Emotive Discourse Amalgamation," in IEEE/ACM Exchanges on Sound, Discourse, and Language Handling, vol. 29, pp. 874-886, 2021, doi: 10.1109/TASLP.2021.3052688.
43. L. Yi and M. - W. Mak, "Further developing Discourse Feeling Acknowledgment With Ill-disposed Information Expansion Organization," in IEEE Exchanges on Brain Organizations and Learning Frameworks, vol. 33, no. 1, pp. 172-184, Jan. 2022, doi: 10.1109/TNNLS.2020.3027600.
44. W. Zhang, P. Tune, D. Chen, C. Sheng and W. Zhang, "Cross-Corpus Discourse Feeling Acknowledgment In view of Joint Exchange Subspace Learning and Relapse," in IEEE Exchanges on Mental and Formative Frameworks, vol. 14, no. 2, pp. 588-598, June 2022, doi: 10.1109/TCDS.2021.3055524.
45. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy and T. H. Falk, "Element Pooling of Tweak Range Highlights for Further developed Discourse Feeling Acknowledgment in the Wild," in IEEE Exchanges on Emotional Figuring, vol. 12, no. 1, pp. 177-188, 1 Jan.- Walk 2021, doi: 10.1109/TAFFC.2018.2858255.
46. W. Zhang and P. Tune, "Move Meager Discriminant Subspace Learning for Cross-Corpus Discourse Feeling Acknowledgment," in IEEE/ACM Exchanges on Sound, Discourse, and Language Handling, vol. 28, pp. 307-318, 2020, doi: 10.1109/TASLP.2019.2955252.

47. Z. Aldeneh and E. M. Executive, "You're Not You When You're Furious: Powerful Inclination Elements Arise by Perceiving Speakers," in IEEE Exchanges on Full of feeling Figuring, vol. 14, no. 2, pp. 1351-1362, 1 April-June 2023, doi: 10.1109/TAFFC.2021.3086050.
48. P. Melody, W. Zheng, Y. Yu and S. Ou, "Discourse Feeling Acknowledgment In light of Powerful Discriminative Meager Relapse," in IEEE Exchanges on Mental and Formative Frameworks, vol. 13, no. 2, pp. 343-353, June 2021, doi: 10.1109/TCDS.2020.2990928.
49. Y. - S. Joo, H. Bae, Y. - I. Kim, H. - Y. Cho and H. - G. Kang, "Viable Feeling Transplantation in a Start to finish Text-to-Discourse Framework," in IEEE Access, vol. 8, pp. 161713-161719, 2020, doi: 10.1109/ACCESS.2020.3021758.
50. M. Ren, X. Huang, X. Shi and W. Nie, "Intuitive Multimodal Consideration Organization for Feeling Acknowledgment in Discussion," in IEEE Signal Handling Letters, vol. 28, pp. 1046-1050, 2021, doi: 10.1109/LSP.2021.3078698.
51. J. Han, Z. Zhang, Z. Ren and B. Schuller, "EmoBed: Fortifying Monomodal Feeling Acknowledgment by means of Preparing with Crossmodal Feeling Embeddings," in IEEE Exchanges on Emotional Registering, vol. 12, no. 3, pp. 553-564, 1 July-Sept. 2021, doi: 10.1109/TAFFC.2019.2928297.
52. S. Hamsa, Y. Iraqi, I. Shahin and N. Werghi, "An Improved Feeling Acknowledgment Calculation Utilizing Pitch Correlogram, Profound Scanty Lattice Portrayal and Arbitrary Backwoods Classifier," in IEEE Access, vol. 9, pp. 87995-88010, 2021, doi: 10.1109/ACCESS.2021.3086062.
53. M. Wu, W. Su, L. Chen, W. Pedrycz and K. Hirota, "Two-Stage Fluffy Combination Based-Convolution Brain Organization for Dynamic Feeling Acknowledgment," in IEEE Exchanges on Emotional Processing, vol. 13, no. 2, pp. 805-817, 1 April-June 2022, doi: 10.1109/TAFFC.2020.2966440.
54. H. Zhao, Y. Xiao and Z. Zhang, "Hearty Semisupervised Generative Antagonistic Organizations for Discourse Feeling Acknowledgment through Conveyance Perfection," in IEEE Access, vol. 8, pp. 106889-106900, 2020, doi: 10.1109/ACCESS.2020.3000751.
55. Nguyen et al., " Profound Auto-Encoders With Consecutive Learning for Multimodal Layered Feeling Acknowledgment," in IEEE Exchanges on Sight and sound, vol. 24, pp. 1313-1324, 2022, doi: 10.1109/TMM.2021.3063612.
56. R. Sadiq and E. Erzin, "Feeling Subordinate Space Transformation for Discourse Driven Emotional Facial Component Amalgamation," in IEEE Exchanges on Full of feeling Registering, vol. 13, no. 3, pp. 1501-1513, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.3008456.
57. Hwang and J. - H. Chang, "Start to finish Discourse Endpoint Discovery Using Acoustic and Language Displaying Information for Online Low-Dormancy Discourse Acknowledgment," in IEEE Access, vol. 8, pp. 161109-161123, 2020, doi: 10.1109/ACCESS.2020.3020696.
58. Du, S. Long and H. Yuan, "Non-Contact Feeling Acknowledgment Joining Pulse and Look for Intuitive Gaming Conditions," in IEEE Access, vol. 8, pp. 11896-11906, 2020, doi: 10.1109/ACCESS.2020.2964794.
59. L. Yang and S. - F. Qin, "A Survey of Feeling Acknowledgment Strategies From Keystroke, Mouse, and Touchscreen Elements," in IEEE Access, vol. 9, pp. 162197-162213, 2021, doi: 10.1109/ACCESS.2021.3132233.
60. M. Tahon, G. Lecorvé and D. Lolive, "Might We at any point Produce Profound Articulations for Expressive Discourse Union?," in IEEE Exchanges on Emotional Processing, vol. 11, no. 4, pp. 684-695, 1 Oct.- Dec. 2020, doi: 10.1109/TAFFC.2018.2828429.